

标准文献知识服务系统设计与实现*

丁　恒[1,2]　陆　伟[2]
[1](国家领土主权与海洋权益协同创新中心　武汉　430072)
[2](武汉大学信息管理学院　武汉　430072)

摘要：【目的】建设面向知识层次的标准文献服务系统，推进标准文献信息服务的知识化进程。【应用背景】标准文献知识服务系统能够对标准文献中的知识单元进行语义抽取，依据标准文献知识之间的关联关系进行有效组织，并为用户提供面向知识层次的标准文献信息服务。【方法】采用光符识别、自然语言处理、信息可视化等技术实现标准文献的语义组织、知识抽取、本体构建、知识图谱、本体检索等功能。【结果】用户利用标准文献知识服务系统，能够获得面向知识层次的标准文献信息服务，包括标准知识图谱和基于本体的标准知识检索服务。【结论】标准文献知识服务系统能够改善用户体验，满足用户的标准文献知识需求。

关键词：标准文献　知识服务　知识组织　本体检索　知识图谱
分类号：G350

1　引　言

信息技术的发展改变了社会信息的传播方式，这种变化对情报学研究内容、方法乃至于对象提出了新的要求。从信息的具体呈现形式看，多样化的数字信息资源丰富了情报学的研究对象，情报研究不再局限于学术文献，网页、图书、专利、档案、标准文献等信息载体都开始为情报研究者所关注[1-3]；从信息链[4]或情报学基本概念[5]的角度思考，情报学研究的对象从信息层次向知识层次深入，如何实现对各种信息载体中知识的有序组织，并提供有效的知识服务成为情报学研究的重要前沿。

标准文献作为一种重要的信息来源和知识载体，其在数字网络环境下的生产、组织、利用也面临着向知识服务方向发展的问题[6]。然而当前的标准文献服务系统多处于文献粒度[7-9]，无法满足用户的知识需求。基于上述思考，本文结合光符识别、自然语言处理、信息可视化、信息检索等技术，构建了面向知识层次的标准文献信息服务系统，为用户提供诸如知识抽取、知识图谱、知识搜索等标准知识服务，在一定程度上提高了标准文献信息服务的质量，改善了用户体验。

2　现状分析

1984年，袁锁鸿[10-11]指出标准情报资料的“反求工程”是提高我国产品质量和国际竞争力的重要手段，并提出了用情报收集、整理、编目和利用的思想为国家、企业、社会提供标准情报服务。20世纪90年代，楼青[12-13]、孙秉秀[14]都从标准情报资料的管理和咨询服务方面，讨论了如何优化标准文献服务工作。由此看来，很早以前标准文献服务就已经是图书情报学界关注的研究问题之一。2000年后，随着计算机技术的普及，国内图书情报学界开始对标准文献的检索服务进行了相关探讨。夏巨岚等[15]、邓要武[8]先后对国内外标准文献信息资源的建设及国内外标准文献检索和服务平台进行了系统调研和比较，结果显示当时国内外

通讯作者：陆伟，ORCID：0000-0002-0929-7416，E-mail：weilu@whu.edu.cn。
*本文系国家领土主权与海洋权益协同创新中心和武汉大学信息管理学院合作项目“边海问题舆情系统开发”的研究成果之一。

标准文献检索服务都存在检索字段较少，文献加工粒度较粗，无法揭示标准文献之间的关联等缺陷，标准文献服务系统研究进入瓶颈期。2011 年，刘佳等[7]对国际标准化组织(ISO)、国际电工委员会(IEC)、国际电信联盟(ITU)、欧洲标准化委员会(CEN)等 4 个标准检索文献平台进行详细调研，结果显示当前的国际标准文献检索服务系统主要包括"标准编号"、"标准名称"、"摘要"、"主题词"、"ICS 号"、"时间范围"、"标准状态"等检索字段，说明标准文献服务系统建设还处于"基于关键字的检索"阶段，没有能够深入标准文献内部的知识单元，也不能够为用户提供面向知识层次的标准文献服务。

21 世纪，语义技术的发展为信息加工和信息服务领域带来新的变革。2001 年，Berners-Lee 和 Hendler[16]指出语义技术将改变知识的生产和分享模式，拉开了基于语义技术和本体的知识服务研究序幕。如 Alani 等[17]提出了一种基于本体的、自动的知识抽取方法，并将其用于网页知识的抽取和标注，推动了网页知识服务的发展。Ghoula 等[18]则利用知识本体对专利文献信息进行语义标注，并基于结构化的专利语义文档构建了专利知识检索和挖掘系统。Muller 等[19]针对生物医学学术文献，构建了包含 33 类术语的医学知识本体，并利用知识本体中术语之间的关系构建语义查询，实现了基于本体的生物医学文献知识检索。此外越来越多的出版商开始借助 XML 语言组织文档，利用丰富的语义标签对科学文献进行语义增强和知识组织[20-21]，为科学文献知识服务提供了崭新思路。这些研究都利用了语义化技术和本体作为领域知识组织的媒介和桥梁，一定程度上反映了学界对知识服务技术手段和实现途径的共识。本体和语义技术在众多领域内的成功应用，给国内标准文献的知识加工和知识服务研究带来了启示。计雄飞等[22]对国内外标准文献专题研究的问题进行分析，指出开展多种服务方式、集成大量信息、进行深层次知识组织和挖掘的标准文献服务的必要性，同时提出了利用数据语义深加工和主题词表等手段开展标准文献知识服务的前瞻意见。李景等[23]尝试以人工语料编辑的形式(概念词汇及其关系)，通过构建标准文献语料库的原型系统，实现了标准文献的分专业领域浏览、双语模糊检索、词汇语义拓展检索等功能，然而这种基于人工语料编辑的知识服务系统要求大量的人工成本，较弱的自动化程度阻碍了该系统的有效应用。

总的来说，当前的标准文献服务系统还停留在基于关键字的文献检索层次，标准文献的加工粒度较粗，未能深入到标准文献内部的语义知识单元，忽视了标准文献知识单元之间的关联关系，因此亟需对标准文献内容进行语义组织和知识抽取，从文献服务系统向知识服务系统转变。

3 系统设计

3.1 问题与思路

通过文献调研和实践调查，笔者发现构建面向知识层次的标准文献知识服务系统主要面临以下问题:

(1) 语义数据缺失。由于标准文献的版权问题导致了标准文献机器可读全文数据获取困难[24]，当前大多数标准文献以 PDF 图像扫描件的形式存储，导致了计算机难以直接读取标准文献内容信息。因此，多数系统通过光符识别技术对标准文献进行转化处理，获取相应纯文本数据。然而，纯文本数据在解决内容读取问题的同时，也导致了原始文本的结构信息丢失，不利于标准文献的语义加工以及语义知识组织。

(2) 领域本体复杂，人工构建困难。由于标准文献覆盖不同专业、行业，涉及众多学科知识内容，本体内容较为复杂，难以构建一个通用的知识本体，并且人工本体构建的方式需要消耗大量的人力成本。

鉴于以上问题，笔者认为构建标准文献知识服务系统的核心任务主要有"标准文献数据的语义再结构化"、"标准领域知识本体的自动构建"。其实现的具体思路如下:

(1) 语义组织：通过图像处理和语义抽取模块对标准文献(PDF或图像格式)进行语义结构重构，将其转化为具有丰富语义结构信息的XML文件。

(2) 知识抽取：利用自然语言处理技术抽取标准文献XML文件中的重要概念。鉴于标准文献的内容往往是对标准化对象的强制性要求或指导性建议，标准文献的知识关联往往以标准化对象为线索，因此系统主要抽取代表标准化对象的概念词汇。

(3) 本体构建：通过外部资源(标准术语文件、网络百科等)抽取与标准化对象相关的语义文本描述，利用语义处理技术抽取相关实体概念，初步自动构

建标准实体之间相互关系。并提供一个人工辅助编辑的接口，允许专家对自动生成的标准知识本体进行人工矫正。

(4) 知识图谱和本体检索：在语义组织和本体构建的基础上，利用可视化技术以知识图谱的形式展示标准知识联系，且通过标准知识本体中的概念关系对基于关键字的标准检索服务进行查询拓展。

3.2 系统软件架构

根据系统设计思路与SOA架构进行系统设计，本系统软件架构主要分为三个层次，如图1所示：

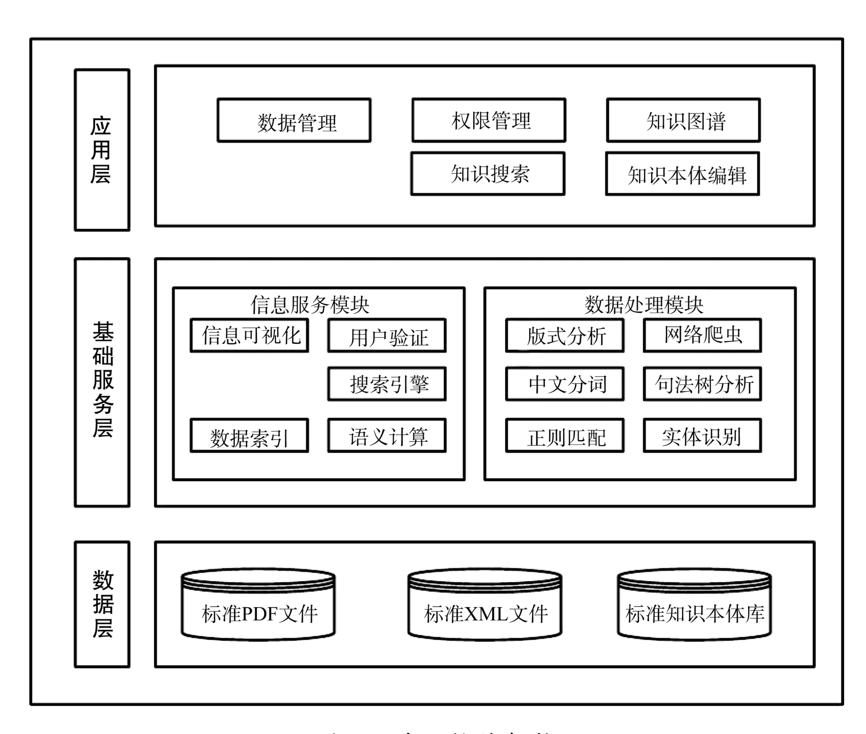


图1　系统软件架构

(1) 数据层：用于执行标准PDF文件、标准文献XML文件及标准知识本体数据库的持久化操作。

(2) 基础服务层：包括数据处理与信息服务两大模块。数据处理模块首先利用版式分析和正则匹配技术将标准PDF文件转化为具有语义结构的标准文献XML文件；其次使用中文分词和实体识别技术从标准XML文件中抽取标准实体对象，并利用网络爬虫技术采集与标准化对象相关的文本描述，最后使用句法树分析初步自动构建标准实体之间的相互关系。同时，信息服务模块利用信息可视化技术为知识图谱提供技术支撑，语义计算、数据索引、搜索引擎和标准知识本体库则是知识搜索服务中的基本单元。

(3) 应用层：包括数据管理、权限管理、知识图谱、知识搜索、知识本体编辑等。

4 核心功能实现

4.1 标准文献数据的语义再结构化

如前文所述，标准知识服务系统的核心步骤和功能是标准文献语义化处理，即将原始的标准PDF文件转化为具有语义结构信息的标准文献XML文件。标准文献的语义化处理本质是指对数字标准文献进行语义加工，使得标准文献的内容片断包含语义标签，将标准文献中的信息知识表示成计算机可读、可识别、可处理的形式，从而使得标准文献的组织方式从文献粒度的树形分类结构向知识粒度的网络结构转变，其直接的形式表现为“借助XML语言为标准文献提供可操作性原始数据”。标准文献数据的语义再结构化流程如图2所示。

管理员用户可通过网页客户端界面上传标准PDF

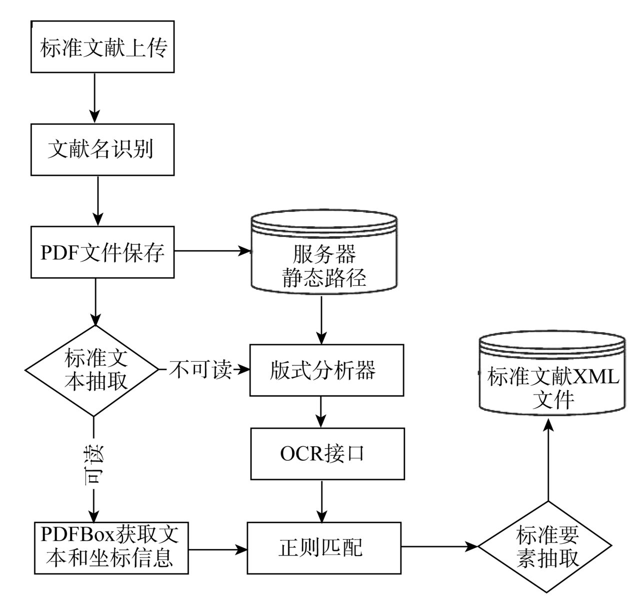


图 2　标准文献数据的语义化再结构化流程

文件数据，系统将该 PDF 文件保存到服务器的固定静态路径，便于其他用户访问。然后系统将判断 PDF 文件格式，对于机器可读的 PDF 文件，使用 PDFBox 工具包[①]获取 PDF 的文本内容和相应坐标信息。若 PDF 文件为图片扫描件，则使用版式分析器对图片进行结构分析，提取图片中的文本区块坐标，并利用光符识别接口(tesseract[②])识别文本区块中的文本内容。版式分析的关键代码如下所示：

```
// 版式分析器
Void text_detect(Mat& image, vect<Rect>& text_regions, int
  char_space){
  // 灰度化处理
  Mat gray_image;
  // 获取 MSER 特征描述符
  MSER mserExtractor;
  mserExtractor = MSER::MSER();
  ……
  mserExtractor(blur_image, regions, Mat());
  // 根据规则获取候选文字区域
  Vector<vector<Point>> canidate_regions;
  remove_regions_by_rule(regions, canidate_regions);
  // 按从上到下从左到右顺序对区域进行排序
  sort_text_rects(mask_rects, canidate_regions);
}
```

利用正则匹配对标准文献的通用要素进行抽取，标准文献的通用要素语义标签如表 1 所示：

表 1　数字标准文献的语义标签列表

语义标签	含义
document_number	标准编号
document_name_in_chinese	中文标题
document_name_in_english	英文标题
date_of_announcement	实施日期
publish_date	发布日期
Orgnization	发布组织
cites_standard	引用标准
Terms	术语

正则匹配的规则算法如下所示：

```
// region (通过版式分析器获取的文本区块)
// region_text (通过 OCR 接口获取的文本区块对应的文本内容
  信息)
// page (PDF 文件的页码信息)
// StandardCode (中标分类号，如 GB、FZ 等)
For regions in page 1: // 如果是 PDF 第一页
  // 如果文本区块包含中标分类号和数字
  If region_text startwith StandardCode and contains numbers:
    Label the region with document_number // 标记为标准编号
  If characters of region_text are all english alphabets:
    Label the region with document_name_in_english
    And Label the previous region with document_name_in__
      chinese
  ……
  If region_text startwith numbers and contains “引用文件”:
    The next regions may be cites_standards
    For region in next regions:
      If text of region startwith StandardCode:
        Label the region with cites_standard
  ……
```

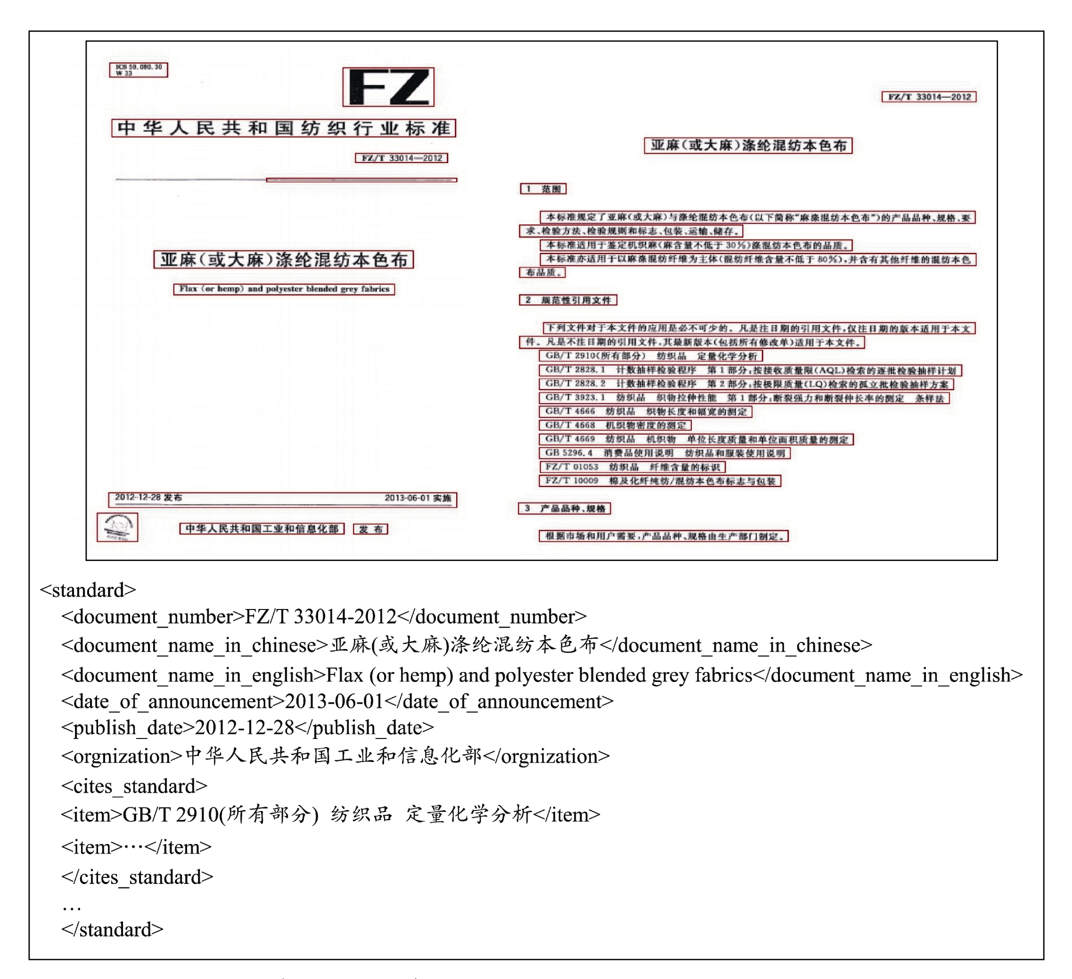
最终的标准文献数据的语义再结构化处理结果以 XML 文件形式保存，结果样例如图 3 所示。

4.2　标准领域知识本体自动构建

标准领域知识本体自动构建的实质是利用中文分词、词性标注、句法分析、序列标注等自然语言处理技术和概率模型对标准文献 XML 文件进行深度加工的过程，最终将标准实体与实体关系构成的语义网络即标准知识本体存入数据库中。其具体流程如图 4 所示。首先，系统从标准文献 XML 数据库中提取某一标准文件的片段信息(如中文标准名)；然后对该片段进

①https://pdfbox.apache.org/.

②https://github.com/tesseract-ocr/tesseract.



FZ

中华人民共和国纺织行业标准

FZ/T 33014—2012

亚麻(或大麻)涤纶混纺本色布

Flax (or hemp) and polyester blended grey fabrics

2012-12-28 发布　　2013-06-01 实施

中华人民共和国工业和信息化部　发布

FZ/T 33014—2012

亚麻(或大麻)涤纶混纺本色布

1　范围

本标准规定了亚麻(或大麻)与涤纶混纺本色布(以下简称"麻涤混纺本色布")的产品品种、规格、要求、检验方法、检验规则和标志、包装、运输、储存。

本标准适用于鉴定机织麻(麻含量不低于30%)涤混纺本色布的品质。

本标准亦适用于以麻涤混纺纤维为主体(混纺纤维含量不低于80%),并含有其他纤维的混纺本色布品质。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是不注日期的引用文件,其最新版本(包括所有修改单)适用于本文件。

GB/T 2910(所有部分)　纺织品　定量化学分析

GB/T 2828.1　计数抽样检验程序　第1部分:按接收质量限(AQL)检索的逐批检验抽样计划

GB/T 2828.2　计数抽样检验程序　第2部分:按极限质量(LQ)检索的孤立批检验抽样方案

GB/T 3923.1　纺织品　织物拉伸性能　第1部分:断裂强力和断裂伸长率的测定　条样法

GB/T 4666　纺织品　织物长度和幅宽的测定

GB/T 4668　机织物密度的测定

GB/T 4669　纺织品　机织物　单位长度质量和单位面积质量的测定

GB 5296.4　消费品使用说明　纺织品和服装使用说明

FZ/T 01053　纺织品　纤维含量的标识

FZ/T 10009　棉及化纤纯纺/混纺本色布标志与包装

3　产品品种、规格

根据市场和用户需要,产品品种、规格由生产部门制定。

<standard>
<document_number>FZ/T 33014-2012</document_number>
<document_name_in_chinese>亚麻(或大麻)涤纶混纺本色布</document_name_in_chinese>
<document_name_in_english>Flax (or hemp) and polyester blended grey fabrics</document_name_in_english>
<date_of_announcement>2013-06-01</date_of_announcement>
<publish_date>2012-12-28</publish_date>
<orgnization>中华人民共和国工业和信息化部</orgnization>
<cites_standard>
<item>GB/T 2910(所有部分) 纺织品 定量化学分析</item>
<item>…</item>
</cites_standard>
…
</standard>

图 3　基于版式分析和正则匹配的标准文献语义化示例图

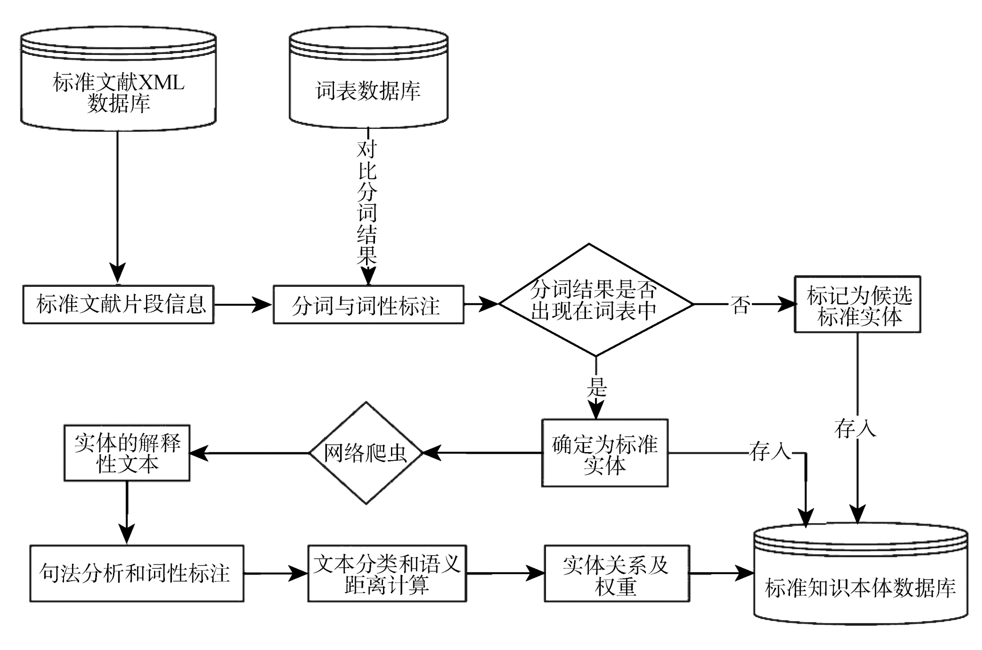


图 4　标准领域知识本体自动构建流程

行中文分词与词性标注(Stanford Parser[①]), 并将分词结果与词表数据库进行对比, 若分词结果存在于现有词表中则将实体存入标准知识本体数据库, 若不在实体词表中则将其标记为候选实体, 并存入标准知识本体数据库; 最后, 对于确定的标准实体对象, 利用网络爬虫获取其相关的解释性文本, 对该文本进行词性标注与句法分析, 抽取其中的名词、动词和句法结构用于文本分类和语义距离计算, 文本分类用于确定两个实体之间的关系, 语义距离计算用于确定两个实体词汇之间的关系权重, 其中文本分类采用文献[25]提出的方法。

实体关系权重计算的形式化过程如下: 对于任意标准实体词汇 E, 通过网络爬虫获取相关解释性文档集合 $D=\{d_1,d_2,\cdots,d_n\}$, d 代表实体词汇 E 一个解释性文档。对每个解释性文档 d_i 进行句子切分得到 $S=\{s_1,s_2,\cdots,s_t\}$, s 代表句子, t 代表句子在文档中的位置顺序。每个句子 s_j 可通过分词、词性标注抽取其中所有的名词和动词 $W=\{w_1,w_2,\cdots,w_k\}$。如果词汇 w_k 在实体词汇表中, 则计算实体词汇 E 与 w_k 之间的关系权重, 记为 $rel(E,w_k)$。该权重使用加权共现算法进行计算, 具体算法如下:

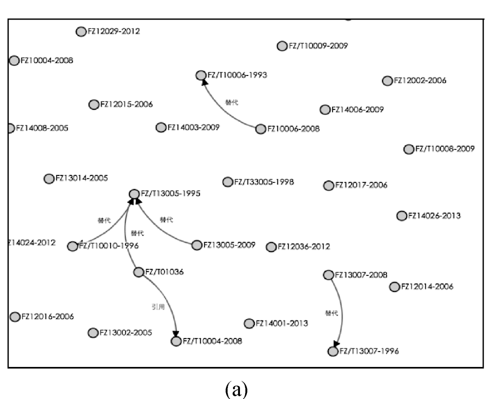
```
rel(E, w_k) = 0
n = len(D)   //总文档数
for d in D:
   co-occurrence = 0   //共现次数
   S = get_sentence(d)   //文档 d 句子切分
   m = len(S)   //文档 d 句子的个数
   for s in S:
       t   //句子 s 在文档中的位置
       W = pos_cut(s)   //分词和词性标注
       W = remove(s)   //去除非名词和动词
       if w_k in W and E in w_k:
             co-occurrence += 1   //共现次数+1
     elif w_k in W and E not in w_k:
            co-occurrence += 1/sqrt(t)
            //共现次数加位置 t 的平方根的倒数
rel(E, w_k) += co-occurrence/m   //每个文档关系权重累加
rel(E, w_k) = rel(E, w_k)/n   //归一化
```

5　系统应用与评价

依据图 1 的软件架构对系统进行开发, 主要向用户提供标准知识图谱服务、基于本体的标准检索服务。系统以 B/S 架构设计, 基础服务模块以后台服务的方式在服务器上运行, 前台为用户提供交互界面。通过与标准文献机构的合作, 获取了包括国家标准、行业标准在内的标准 PDF 文件共 2 268 篇, 运用本系统对这些标准文献进行语义再结构化。

(1) 标准文献知识图谱服务

标准文献知识图谱服务是从不同的信息粒度上解释标准对象之间的关系, 用可视化技术对这些关系进行展示, 辅助用户理解和利用标准文献知识。实际应用中, 系统主要从文献和实体两个粒度上对标准文献知识进行可视化展示。在标准文献粒度层次上, 标准文献之间主要存在着引用和替代两种关系, 这种关系一定程度上揭示了标准知识的更替, 是标准知识利用的重要参考依据。图 5(a)是标准文献粒度上知识关系的可视化图谱, 用户通过查询某一标准产品对象(图 5(a)中以"棉纺"为查询词), 系统为用户返回主题词相关的标准文献, 并提供标准文献之间的引用替代关系。在标准实体粒度层次上, 系统主要展示标准实体之间的联系, 图 5(b)是标准实体"喷胶棉絮片"的知识可视化展示, 用户可清晰地查看与"喷胶棉絮片"相关的实体概念。



(a)

(b)

图 5　标准文献知识图谱服务界面

①http://nlp.stanford.edu/software/lex-parser.shtml.

(2) 基于本体的标准检索服务

目前，基于关键字的标准检索服务没有考虑到标准知识之间的内在联系，如图6(a)是以“喷胶棉絮片”为检索词在国家标准文献共享服务平台[①]查询到的标准文献结果列表，该结果仅返回了包含查询关键字的标准文件，然而根据标准实体“喷胶棉絮片”的知识可视化结果可知(图5(b))，“喷胶棉絮片”的生成过程中涉及到“涤纶短纤维”、“纤网”等产品及“梳理机”这一设备，与这些实体相关的标准文献也包含着用户需要的知识片段。为解决这一问题，本系统利用标准领域知识本体中的概念关系，通过查询拓展的手段实现了基于本体的标准检索服务，图6(b)是以“喷胶棉絮片”为检索词在标准文献知识服务系统中查询到的文献列表。可以发现，本文构建的标准知识服务系统能够准确返回包含“喷胶棉絮片”相关标准实体的标准文献，满足了用户实际生产中的标准知识需求。



(a)

喷胶棉絮片 高级搜索 简单搜索

筛选条件设置

编号	序列号	标准名	类别
1	FZ/T 64003-2011	喷胶棉絮片	工业产品
2	FZ/T 93033-2014	梳棉机	纺织机械与器具
3	SB/T 10169-1993	MS-1016型梳棉机	纺织机械与器具
4	FZ/T 93019-2004	梳棉机用弹性盖板针布	纺织机械与器具
5	FZ/T 93066-2007	梳棉机用齿条盖板针布	纺织机械与器具
6	FZ/T 64041-2014	熔喷纤网非织造粘合衬	工业产品
7	FZ/T 64018-2011	纤网—纱线型缝编非织造布	工业产品
8	FZ/T 50004-2011	涤纶短纤维干热收缩率试验方法	性能测定

首页 上一页 1 2 3 下一页 尾页

(b)

图 6　基于本体的标准检索服务

6 结　语

知识组织与知识服务的理念已经渗透进标准文献信息服务领域，面向知识层次的标准文献服务系统是标准文献信息服务领域的发展方向。国内很多图情研究者从理论上论证了标准知识服务的重要性，然而关于标准文献知识服务系统构建的探讨还比较少。本文从应用角度出发设计了一种面向知识层次的标准文献服务系统，并采用光符识别、自然语言处理、信息可视化等技术实现标准文献的语义结构化组织、本体自动构建等功能，为用户提供了本体检索、知识图谱等知识服务。当然，该系统的设计只是对标准文献知识服务系统研究的初步探讨，在未来的研究中还需进一步对标准文献的知识结构、知识组织形式、知识服务方式进行更深入的研究，改善用户界面，提升用户体验。

①http://www.cssn.net.cn/.

参考文献：

[1] 罗立国, 余翔, 郑婉婷, 等. 专利检索网站比较研究[J]. 情报杂志, 2012, 31(3): 163-167. (Luo Liguo, Yu Xiang, Zheng Wanting, et al. Comparative Study on Patent Retrieval Websites[J]. Journal of Intelligence, 2012, 31(3): 163-167.)

[2] 李智锋, 张李义. 混合动力思想下的图书检索系统研究[J]. 现代图书情报技术, 2012(7-8): 54-58. (Li Zhifeng, Zhang Liyi. Research of Books Retrieval System Under Thinking of Hybrid System [J]. New Technology of Library and Information Service, 2012(7-8): 54-58.)

[3] 马费成, 高静. Web2.0 信息半衰期影响因素实证研究——以社会书签网站为例[J]. 情报理论与实践, 2010, 33(11): 1-6. (Ma Feicheng, Gao Jing. Research on Web2.0 Information Half Life Measurement and Its Impact Factors —— Taking Social Bookmark Website as an Example [J]. Journal of Information Studies: Theory and Application, 2010, 33(11): 1-6.)

[4] 马费成. 情报学发展的历史回顾及前沿课题[J]. 图书情报知识, 2013(2): 4-12. (Ma Feicheng. Historical Review of the Development of Information Science with Proposing Frontier Topics [J]. Document,Informaiton & Knowledge, 2013(2): 4-12.)

[5] 郑彦宁, 化柏林. 数据、信息、知识与情报转化关系的探讨[J]. 情报理论与实践, 2011, 34(7): 1-4. (Zheng Yanning, Hua Bolin. Discussion on Transforming Relationship of Data, Information, Knowledge and Intelligence[J]. Journal of Information Studies: Theory and Application, 2011, 34(7): 1-4.)

[6] 郭德华. 标准文献知识链接服务模式研究[J]. 图书情报工作, 2011, 55(9): 76-79. (Guo Dehua. Study on Knowledge Link Service Mode of Standard Literatures[J]. Library and Information Service, 2011, 55(9): 76-79.)

[7] 刘佳, 钟永恒. 国际标准文献检索平台的比较及启示[J]. 图书馆学研究, 2011(20):60-64. (Liu Jia, Zhong Yongheng. Comparison and Enlightenment of the Retrieval Systems of International Standard Document [J]. Researches in Library Science, 2011(20): 60-64.)

[8] 邓要武. 科技报告、专利文献和标准文献资源检索与利用[J]. 图书馆工作与研究, 2008(7): 71-74. (Deng Yaowu. The Retrieval and Applications of Scientific and Technical Report, Patent Document and Standard Literature Resources [J]. Library Work and Study, 2008(7): 71-74.)

[9] 李伟华, 王通, 顾英. 网络标准文献信息资源的分布及检索[J]. 情报探索, 2010(12): 74-77. (Li Weihua, Wang Tong, Gu Ying. The Retrieval and Applications of Scientific and Technical Report, Patent Document and Standard Literature Resources [J]. Information Research, 2010(12): 74-77.)

[10] 袁锁鸿. 应当重视国外标准情报资料的利用[J]. 图书情报知识, 1984(4): 42-43. (Yuan Suohong. Attach Importance to the Use of Foreign Standard Information [J]. Document, Informaiton and Knowledge, 1984(4): 42-43.)

[11] 袁锁鸿. 重视对国外产品标准样本资料的利用[J]. 图书馆杂志, 1985(4): 44-45. (Yuan Suohong. Pay Attention to the Use of Foreign Product Standard Sample Data [J]. Library Journal, 1985(4): 44-45.)

[12] 楼青. 标准文献的管理与服务[J]. 图书馆杂志, 1991(6): 26-27. (Lou Qing. Management and Service of Standard Literature [J]. Library Journal, 1991(6): 26-27.)

[13] 楼青. 优化标准文献的服务[J]. 图书情报知识, 1996(1): 58-59. (Lou Qing. Optimizing the Service of Standard Literature [J]. Document,Informaiton & Knowledge, 1996(1): 58-59.)

[14] 孙秉秀. 标准文献利用的新问题[J]. 图书馆工作与研究, 1994(2): 49-50. (Sun Bingxiu. New Problems in the Use of Standard Literature [J]. Library Work and Study, 1994(2): 49-50.)

[15] 夏巨岚, 翟煜男. 浅谈标准文献检索[J]. 图书馆建设, 2002(3): 99-100. (Xia Julan, Zhai Yu'nan. Standard Literature Retrieval [J]. Library Development, 2002(3): 99-100.)

[16] Berners-Lee T, Hendler J. Publishing on the Semantic Web [J]. Nature, 2001, 410(6832): 1023-1024.

[17] Alani H, Kim S, Millard D E, et al. Automatic Ontology-based Knowledge Extraction from Web Documents [J]. IEEE Intelligent Systems, 2003, 18(1): 14-21.

[18] Ghoula N, Khelif K, Dieng-Kuntz R. Supporting Patent Mining by Using Ontology-based Semantic Annotations[C]. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, 2007: 435-438.

[19] Muller H-M, Kenny E E, Sternberg P W. Textpresso: An Ontology-based Information Retrieval and Extraction System for Biological Literature [J]. PLoS Biology, 2004, 2(11): e309.

[20] 王晓光, 陈孝禹. 语义出版的概念与形式[J]. 出版发行研究, 2011(11): 54-58. (Wang Xiaoguang, Chen Xiaoyu. Concepts and Forms of Semantic Publishing [J]. Publishing Research, 2011(11): 54-58.)

[21] Iorio A D, Lange C, Dimou A, et al. Semantic Publishing Challenge – Assessing the Quality of Scientific Output by Information Extraction and Interlinking [M]. New York: Springer International Publishing, 2015: 65-80.

[22] 计雄飞, 胡雄伟, 张宝林, 等. 面向服务的标准信息专题知识组织[J]. 标准科学, 2010(8): 27-30. (Ji Xiongfei, Hu Xiongwei, Zhang Baolin, et al. Service-oriented Standard Subject Knowledge Organization [J]. Standard Science, 2010(8): 27-30.)

[23] 李景, 李国鹏, 汪滨, 等. 标准文献语料库构建研究[J]. 图书馆理论与实践, 2013(12): 41-44. (Li Jing, Li Guopeng, Wang Bin, et al. Research on the Construction of Standard Document Corpus [J]. Library Theory and Practice, 2013(12): 41-44.)

[24] 赵美娣. 说说标准文献的获取[EB/OL]. [2015-12-19]. http://blog.sciencenet.cn/blog-69474-944498.html. (Zhao Meidi. Talking about the Acquisition of the Standard Literature [EB/OL]. [2015-12-19]. http://blog.sciencenet.cn/blog-69474-944498.html.)

[25] 董静, 孙乐, 冯元勇, 等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-85, 91. (Dong Jing, Sun Le, Feng Yuanyong, et al. Chinese Automatic Entity Relation Extraction[J]. Journal of Chinese Information Processing, 2007, 21(4): 80-85, 91.)

作者贡献声明:

陆伟,丁恒: 提出系统设计思路、设计系统开发方案;
丁恒: 应用系统开发和测试, 论文起草;
陆伟: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

收稿日期: 2016-04-05
收修改稿日期: 2016-05-15

Building Standard Literature Knowledge Service System

Ding Heng[1, 2]　Lu Wei[2]
[1](Collaborative Innovation Center for Territorial Sovereignty and Maritime Rights, Wuhan 430072, China)
[2](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This sutdy builds a knowledge-oriented standard literature service system, which could generate more knowledge for the users. [**Context**] The proposed system is able to extract semantic knowledge unit from the standard literature, to organize information based on the knowledge relationship, and to provide standard knowledge service to users. [**Methods**] We used the technology of optical character recognition, natural language processing, information visualization to finish the tasks of semantic organization, knowledge extraction, Ontology construction, knowledge map and Ontology-based retrieval of standard literature. [**Results**] The users enjoyed knowledge-oriented standard literature information service, including standard knowledge map and Ontology-based retrieval. [**Conclusions**] The proposed system improves user experience and meet their knowledge demands.

Keywords: Standard literature　Knowledge service　Knowledge organization　Ontology-based information retrieval　Knowledge map